Expectation-Propagation for Summary-Less, Likelihood-Free Inference

Simon Barthelmé (BCCN & TU Berlin, simon.barthelme@bccn-berlin.de)
Nicolas Chopin (CREST-ENSAE, nicolas.chopin@ensae.fr)

Abstract

Many models of interest in the natural and social sciences have no closed-form likelihood function, which means that they cannot be treated using the usual techniques of statistical inference. In the case where such models can be efficiently simulated, Bayesian inference is still possible thanks to the Approximate Bayesian Computation (ABC) algorithm. Although many refinements have since been suggested, the technique suffers from three major shortcomings. First, it requires introducing a vector of "summary statistics", the choice of which is arbitrary and may lead to strong biases. Second, ABC may be excruciatingly slow due to very low acceptance rates. Third, it cannot produce a reliable estimate of the marginal likelihood of the model.

We introduce a technique that solves the first and the third issues, and considerably alleviates the second. We adapt to the likelihood-free context a variational approximation algorithm, Expectation Propagation (Minka, 2001). The resulting algorithm is shown to be faster by a few orders of magnitude than alternative algorithms, while producing an overall approximation error which is typically negligible. Comparisons are performed in three real-world applications which are typical of likelihood-free inference, including one application in neuroscience which is novel, and possibly too challenging for standard ABC techniques.

Key-words: Approximate Bayesian Computation; Expectation Propagation; Likelihood-Free Inference; Quasi-Monte Carlo.

1 Introduction

In natural and social sciences, one finds many examples of probabilistic models whose likelihood function is intractable. This includes most models of noisy biological neural networks (Gerstner and Kistler, 2002), some time series and choice models in Economics (Train, 2003), phylogenetic models in evolutionary Biology (Beaumont, 2010), among others. That the likelihood is intractable is unfortunate, because one would still like to perform the usual statistical tasks of parameter inference and model comparison, and the traditional statistical tool-kit assumes that the likelihood function is either directly available or can be made so by introducing latent variables. This explains that researchers have often had to content themselves with semi-quantitative analyses showing that a model could reproduce some aspect of an empirical phenomenon for some values of the parameters; for two representative examples from vision science, see Nuthmann et al., 2010; Brascamp et al., 2006.

A breakthrough was provided by the work of Pritchard et al. (1999) and Beaumont et al. (2002), in the form of the Approximate Bayesian Computation algorithm, which enables Bayesian inference in the likelihood-free context. Assuming some model $p(\mathbf{y}^*|\boldsymbol{\theta})$ for the data $\boldsymbol{y}^\star$, a prior $p(\boldsymbol{\theta})$ over the parameter $\boldsymbol{\theta} \in \Theta$, the following procedure produces exact samples from the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$:

1. Draw $\boldsymbol{\theta}$ from the prior, $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$.
2. Draw a dataset $\mathbf{y}$ from the model conditional on $\boldsymbol{\theta}$, $\boldsymbol{y}|\boldsymbol{\theta} \sim p(\boldsymbol{y}|\boldsymbol{\theta})$.
3. If $\mathbf{y} = \mathbf{y}^*$ then keep $\boldsymbol{\theta}$, otherwise reject.

The acceptance rate provides an estimate of the evidence (marginal likelihood) $p(\boldsymbol{y}^\star) = \int p(\boldsymbol{\theta})p(\boldsymbol{y}^\star|\boldsymbol{\theta})\, d\boldsymbol{\theta}$. Of course, if $\mathbf{y}$ is continuous, then the procedure is hopeless - we will end up rejecting every value of $\boldsymbol{\theta}$ we draw. Even if $\mathbf{y}$ is discrete, the acceptance probability may still be ridiculously low. The solution put forward by Pritchard et al. (1999) is to define some pseudo-distance $d(\mathbf{y}, \mathbf{y}^*)$ between real and simulated datasets and accept samples if the distance is smaller than some value ϵ. This produces samples from the so-called ABC posterior:

$$p_\epsilon\left(\boldsymbol{\theta}|\mathbf{y}^*\right) \propto p(\boldsymbol{\theta}) \int p(\mathbf{y}|\boldsymbol{\theta}) \mathbb{1}_{\{d(\mathbf{y},\mathbf{y}^\star)\leq\epsilon\}}\, d\mathbf{y}. \tag{1.1}$$

The pseudo-distance is usually taken to be $d(\boldsymbol{y}, \boldsymbol{y}^\star) = ||\mathbf{s}\left(\mathbf{y}\right) - \mathbf{s}\left(\mathbf{y}^*\right)||$, for some norm $||\cdot||$, where $\mathbf{s}(\mathbf{y})$ is a vector of summary statistics, for example some empirical quantiles or moments of $\mathbf{y}$. Unless $\boldsymbol{s}$ is sufficient, the approximation

error does not vanish as $\epsilon \to 0$, i.e. $p(\boldsymbol{\theta}|\boldsymbol{s}(\mathbf{y}^\star)) \neq p(\boldsymbol{\theta}|\mathbf{y}^\star)$. In that respect, the ABC posterior suffers from two levels of approximation: a nonparametric error governed by the bandwidth ϵ (see e.g. Blum, 2010), and a bias introduced by the summary statistics $\mathbf{s}$. The more we include in $\mathbf{s}(\mathbf{y})$, the smaller the bias induced by $\mathbf{s}$ should be. On the other hand, as the dimensionality of $\mathbf{s}(\mathbf{y})$ increases, the lower the acceptance rate will be. We would then have to increase ϵ, which leads to an approximation of lower quality.

Thus, ABC requires in practice some more or less arbitrary compromise between what summary statistics to include and how to set ϵ. To establish that the results of the inference are somewhat robust to these choices, many runs of the algorithm are required. Although several variants of the original ABC algorithm that aim at increasing acceptance rates exist (e.g. Beaumont et al., 2002), the current state of the matter is that an ABC analysis is very far from routine use because it may take days to tune on real problems.

Another important limitation of ABC is that introducing a summary statistics $\boldsymbol{s}$ makes it impossible to perform model choice in most cases (Robert et al., 2011). The normalising constant of (1.1) approximates $p(\boldsymbol{s}(\boldsymbol{y}^\star))$, not $p(\boldsymbol{y})$, and, except in very special cases, the former quantity has no clear interpretation in terms of model choice.

In this article we introduce EP-ABC, an adaptation of the Expectation Propagation algorithm (Minka, 2001; Bishop, 2006, Chap. 10) to the likelihood-free setting. EP-ABC is much faster than previous ABC algorithms: typically, it provides accurate results in a few minutes, whereas a standard ABC algorithm need a few hours. EP-ABC incorporates one data-point y_i at a time, which makes it possible to do away with summary statistics, and constrain all the data-points y_i to be close to the $y_i^\star$'s, while maintaining a reasonable computational cost. More formally, EP-ABC builds an EP approximation of the following type of ABC posterior distributions:

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}) \mathbb{1}_{\left\{\|y_i - y_i^\star\| \leq \varepsilon\right\}} \, dy_i \right\} \tag{1.2}$$

with the convention that $p(y_1|y_{1:0}^\star, \boldsymbol{\theta}) = p(y_1|\boldsymbol{\theta})$, and assuming that the datasets $\mathbf{y}$ and $\mathbf{y}^\star$ may be decomposed into n random variables, y_i and $y_i^\star$. When $\epsilon \to 0$, this ABC posterior converges to the true posterior, and its normalising constant (up to to some simple transformation we describe later) converges to the the true evidence $p(\boldsymbol{y}^\star)$.

EP-ABC also suffers from two levels of approximation (EP, then ABC). However, as we show in the cases we study, one is better off with a good approximation to a very good ABC approximation to the posterior (that is, corresponding to a small value of ϵ, and not based on a summary statistics), than with an exact version of a bad ABC approximation (based on a large ϵ and some arbitrary set of summary statistics).

We start with a generic description of the EP algorithm, in Section 2, and explain in Section 3 how it can be adapted to the likelihood-free setting. We explain in Section 4 how EP-ABC can be made particularly efficient when data-points are IID (independent and identically distributed). Section 5 contains three case studies drawn from Finance, population ecology, and vision science. The two first examples are borrowed from already known applications of ABC, and illustrate to which extent EP-ABC may outperform standard ABC techniques in realistic scenarios. To the best of our knowledge, our third example from vision science is a novel application of likelihood-free inference, which, as which shall argue, seems too challenging for standard ABC techniques.

We use the following notations throughout the paper: bold letters refer to vectors or matrices, e.g. $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}$ and so on. We also use bold face to distinguish complete sets of observations, i.e. $\boldsymbol{y}$ or $\boldsymbol{y}^\star$, from their components, y_i, and $y_i^\star$, $i = 1, \dots n$, although we do not assume that these components are necessarily scalar. For sub-vectors of observations, we use the colon notation: $y_{1:i} = (y_1, \dots, y_i)$. The notation $\|\cdot\|$ refers to a generic norm, and $\|\cdot\|_2$ refers to the Euclidean norm. The Kullback-Leibler divergence between probability densities π and q is denoted as

$$KL(\pi\|q) = \int \pi(\boldsymbol{\theta}) \log\left(\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}.$$

The letter p always refers to the probability densities concerning the model; i.e. $p(\boldsymbol{\theta})$ is the prior, $p(y_1|\boldsymbol{\theta})$ is the likelihood of the first observation, and so on. Transpose of a matrix $\boldsymbol{A}$ is denoted $\boldsymbol{A}^t$.

2 Expectation Propagation

Expectation Propagation (EP, Minka, 2001) is an algorithm for variational inference, a class of techniques that aim at finding a tractable probability distribution $q(\boldsymbol{\theta})$ that best approximates an intractable target density $\pi(\boldsymbol{\theta})$. One way to formulate this goal is as finding the member of some parametric family $\mathcal{Q}$ that is in some sense closest to $\pi(\boldsymbol{\theta})$, where "closest" is defined by some divergence between probability distributions. Many variational methods (e.g. Variational Bayes, see Chap. 10 of Bishop, 2006) have as goal the minimisation of the Kullback-Leibler divergence between q and π, $KL(q||\pi)$. This leads to very fast algorithms, but the resulting approximation has some unfortunate properties, for

example that of systematically underestimating the variance (Wang and Titterington, 2005). EP tries instead to minimise $KL(\pi||q)$, which may be a more sensible goal from a statistical point of view.

We begin our discussion of EP with some background material on exponential families and KL divergence.

2.1 Exponential families, KL divergence and the computation of moments

In this paper and in most applications of EP, the family $\mathcal{Q}$ of potential approximations is a particular natural exponential family, with respect to variable $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$:

$$q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \exp\left\{\boldsymbol{\lambda}^t \boldsymbol{t}\left(\boldsymbol{\theta}\right) - \phi\left(\boldsymbol{\lambda}\right)\right\} \tag{2.1}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{d'}$ is the *natural parameter*, $\boldsymbol{t}\left(\boldsymbol{\theta}\right)$ is a *feature map* from Θ to the feature space $\mathbb{R}^{d'}$, and ϕ is known variously as the *partition function* or the *log-cumulant function*. The most interesting property of ϕ is the following:

$$\frac{d}{d\boldsymbol{\lambda}}\phi\left(\boldsymbol{\lambda}\right) = \int \boldsymbol{t}(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\lambda}^t \boldsymbol{t}\left(\boldsymbol{\theta}\right) - \phi(\boldsymbol{\lambda})\right\}\, d\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\lambda}}\left\{\boldsymbol{t}(\boldsymbol{\theta})\right\} \tag{2.2}$$

where $\mathbb{E}_{\boldsymbol{\lambda}}$ denotes an expectation with respect to $q_{\boldsymbol{\lambda}}$. The derivative of the partition function gives the *moment parameters*, usually denoted by $\boldsymbol{\eta}$. It can be shown that $\frac{d}{d\boldsymbol{\lambda}}\phi\left(\boldsymbol{\lambda}\right)$ is, as a function of $\boldsymbol{\lambda}$, a smooth invertible map. Thus, one may also parametrise $q_{\boldsymbol{\lambda}}$ in terms of $\boldsymbol{\eta}$.

In this paper, we will take $\mathcal{Q}$ to be the set of multivariate Gaussian distributions of dimension d. In natural exponential family form, it reads:

$$\boldsymbol{\lambda} = (\boldsymbol{r}, \boldsymbol{Q}), \quad q_{\boldsymbol{\lambda}}\left(\boldsymbol{\theta}\right) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q}\boldsymbol{\theta} + \mathbf{r}^t\boldsymbol{\theta}\right) \tag{2.3}$$

where $\mathbf{Q}$ and $\mathbf{r}$ are the natural parameters. The more familiar moment parameters are given by:

$$\boldsymbol{\eta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{\lambda}}\left(\boldsymbol{\theta}\right) = \mathbf{Q}^{-1}\mathbf{r}, \quad \boldsymbol{\Sigma} = \mathbb{C}\mathrm{ov}_{\boldsymbol{\lambda}}\left(\boldsymbol{\theta}\right) = \mathbf{Q}^{-1},$$

where $\mathbb{C}\mathrm{ov}_{\boldsymbol{\lambda}}$ denotes the Covariance matrix with respect to $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$. To approximate a distribution π with a distribution $q_{\boldsymbol{\lambda}}$ in an exponential family, one may minimise $KL(\pi||q_{\boldsymbol{\lambda}})$ with respect to $\boldsymbol{\lambda}$, which leads to the equation:

$$\begin{aligned}
\frac{d}{d\boldsymbol{\lambda}}KL(\pi||q_\lambda) &= \frac{d}{d\boldsymbol{\lambda}}\left\{\int \pi\left(\boldsymbol{\theta}\right)\log\frac{\pi(\boldsymbol{\theta})}{q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})}\, d\boldsymbol{\theta}\right\} \\
&= \frac{d}{d\boldsymbol{\lambda}}\phi\left(\boldsymbol{\lambda}\right) - \int \pi(\boldsymbol{\theta})\boldsymbol{t}(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 0.
\end{aligned}$$

The solution is therefore obtained by "moment matching": the closest member $q_{\boldsymbol{\lambda}}$ to π is obtained by imposing that the moments of $\boldsymbol{t}(\boldsymbol{\theta})$ under π and under $q_{\boldsymbol{\lambda}}$ are identical. In the Gaussian case, the multivariate Gaussian distribution $N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ closest to π (as measured by KL divergence) is defined by the equations:

$$\boldsymbol{\mu} = \int \boldsymbol{\theta}\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \quad \boldsymbol{\Sigma} = \int (\boldsymbol{\theta} - \boldsymbol{\mu})(\boldsymbol{\theta} - \boldsymbol{\mu})^t \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

2.2 Assumptions of Expectation Propagation

EP assumes that the target density $\pi(\boldsymbol{\theta})$ decomposes into a product of simpler factors, and exploits this factorisation in order to construct a sequence of simpler problems. When $\pi(\boldsymbol{\theta})$ is a posterior density, this factorisation typically takes the following form:

$$\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p\left(\boldsymbol{\theta}\right)\prod_{i=1}^{n} l_i\left(\boldsymbol{\theta}\right) \tag{2.4}$$

where $p(\boldsymbol{\theta})$ is the prior density of parameter $\boldsymbol{\theta}$, and the l_i's correspond to a chain rule decomposition of likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$, e.g. $l_i(\boldsymbol{\theta}) = p(y_i|y_{1:i-1}, \boldsymbol{\theta})$ if $\boldsymbol{y}$ is a set of n observations y_i, with the convention that $p(y_1|y_{1:0}, \boldsymbol{\theta}) = p(y_1|\boldsymbol{\theta})$.

EP uses an approximating distribution with a similar structure:

$$q(\boldsymbol{\theta}) \propto \prod_{i=0}^{n} f_i(\boldsymbol{\theta}) \tag{2.5}$$

where the f_i's are known as the "sites". To obtain a Gaussian approximation, one takes $f_i(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q_i} \boldsymbol{\theta} + \mathbf{r}_i^t \boldsymbol{\theta}\right)$, so that:

$$q(\boldsymbol{\theta}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\theta}^t \left(\sum_{i=0}^{n} \mathbf{Q}_i\right) \boldsymbol{\theta} + \left(\sum_{i=0}^{n} \mathbf{r}_i\right)^t \boldsymbol{\theta} \right\} \tag{2.6}$$

where $\mathbf{Q}_i$ and $\mathbf{r}_i$ are called the *site parameters*. For the sake of simplicity, we assume from now on that the true prior $p(\theta)$ is Gaussian, with natural parameters $\boldsymbol{Q}_0$ and $\boldsymbol{r}_0$. In that case, the site f_0 is kept equal to the prior, and only the sites f_1 to f_n need to be updated. Note however that EP may accommodate more general priors.

In spirit, EP is close to an coordinate-descent optimization algorithm: the sites are updated one by one, in order to progressively minimise the pseudo-distance $KL(\pi \| q)$. The next section describes the site update.

2.3 Site update

Suppose that (2.5) is the current approximation, and one wishes to update site i. This is done by creating a "hybrid" distribution, obtained by substituting site i with the true likelihood contribution $l_i(\boldsymbol{\theta})$:

$$h(\boldsymbol{\theta}) \propto q_{-i}(\boldsymbol{\theta})\, l_i(\boldsymbol{\theta}), \quad q_{-i}(\boldsymbol{\theta}) = \prod_{j \neq i} f_j(\boldsymbol{\theta}). \tag{2.7}$$

For Gaussian sites, this leads to:

$$h_i(\boldsymbol{\theta}) \propto l_i(\boldsymbol{\theta}) \exp\left(-\frac{1}{2}\boldsymbol{\theta}^t \boldsymbol{Q}_{-i} \boldsymbol{\theta} + \mathbf{r}_{-i}^t \boldsymbol{\theta}\right), \quad \boldsymbol{Q}_{-i} = \sum_{j \neq i} \mathbf{Q}_j, \quad \boldsymbol{r}_{-i} = \sum_{j \neq i} \boldsymbol{r}_i.$$

This hybrid distribution may be interpreted as a posterior distribution, based on a Gaussian prior, with natural parameters $\mathbf{Q}_{-i}$ and $\mathbf{r}_{-i}$, and the likelihood of one data-point. The idea is that, by getting closer to the hybrid we should also get closer to the actual density. To find the Kullback-Leibler projection of the hybrid onto $\mathcal{Q}$, we compute its moments, as explained in Section 2.1. In the Gaussian case, this gives:

$$\begin{aligned} Z_h &= \int q_{-i}(\boldsymbol{\theta})\, l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \\ \boldsymbol{\mu}_h &= \frac{1}{Z_h} \int \boldsymbol{\theta} q_{-i}(\boldsymbol{\theta})\, l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \\ \boldsymbol{\Sigma}_h &= \frac{1}{Z_h} \int \boldsymbol{\theta}\boldsymbol{\theta}^t q_{-i}(\boldsymbol{\theta})\, l_i(\boldsymbol{\theta})\, d\boldsymbol{\theta} - \boldsymbol{\mu}_h \boldsymbol{\mu}_h^t. \end{aligned} \tag{2.8}$$

Many models, e.g. Gaussian mixtures, or generalised linear models, are such that these moments admit a closed-form expression, or at least simplify to one-dimensional integrals, which are easy to compute numerically. For our purpose, suffice it to say that the applicability of EP is directly determined by the tractability of the integrals above, that is, how easy it is to compute the two first moments of a pseudo-posterior, consisting of a Gaussian pseudo-prior q_{-i} and one likelihood factor l_i.

Once the moments have been computed, the final step is simply to update the approximation so that it has the same moments as the hybrid, completing the Kullback-Leibler projection. This is done by updating the site parameters for site i: $\boldsymbol{\lambda}_i \leftarrow \boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{-i}$, with $\boldsymbol{\lambda}_h = \boldsymbol{\lambda}(\boldsymbol{\eta}_h)$, and $\boldsymbol{\eta}_h$ is the moment parameters obtained from the hybrid. In the Gaussian case, this gives:

$$\boldsymbol{Q}_i \leftarrow \boldsymbol{\Sigma}_h^{-1} - \boldsymbol{Q}_{-i}, \quad \boldsymbol{r}_i \leftarrow \boldsymbol{\Sigma}_h^{-1} \boldsymbol{\mu}_h - \boldsymbol{r}_{-i}.$$

EP proceeds by looping over sites, updating each one in turn until convergence is achieved. The more general EP algorithm for a generic exponential family is described as Algorithm 1. For the particular case where the exponential family is the family of Gaussian distributions of dimension d, as described above, one simply takes $\boldsymbol{\lambda} = (\boldsymbol{r}, \boldsymbol{Q})$, and $\boldsymbol{\eta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In particular, Step 2 of Algorithm 1 corresponds exactly to computing the moments in (2.8).

2.4 Approximation of evidence

EP also provides an approximation of the normalising constant of (2.4), that is, the evidence $p(\boldsymbol{y})$. This is based on the same ideas of updating site approximations through moment matching. We rewrite in a normalised form the target distribution

Algorithm 1 Generic EP for exponential families.

Input: a target density $\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i=1}^{n} l_i(\boldsymbol{\theta})$.
Initialise $\boldsymbol{\lambda}_0$ to the exponential parameters of the prior p_0, and local site parameters $\boldsymbol{\lambda}_1 \dots \boldsymbol{\lambda}_n = 0$. Set global approximation parameter $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda} = \sum_{i=0}^{n} \boldsymbol{\lambda}_i = \boldsymbol{\lambda}_0$. Loop over sites $i = 1, \dots, n$ until convergence:

1. Create hybrid distribution $h(\boldsymbol{\theta}) \propto q_{-i}(\boldsymbol{\theta}) l_i(\boldsymbol{\theta})$ by setting $q_{-i}(\boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\lambda}_{-i}^{t} \mathbf{t}(\boldsymbol{\theta})\right)$ with $\boldsymbol{\lambda}_{-i} = \boldsymbol{\lambda} - \boldsymbol{\lambda}_i$.
2. Compute moments $\boldsymbol{\eta}_h$ of hybrid distribution, transform to natural parameters $\boldsymbol{\lambda}_h = \boldsymbol{\lambda}(\boldsymbol{\eta}_h)$.
3. Update site i by setting $\boldsymbol{\lambda}_i = \boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{-i}$, then reset global parameter $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda} = \boldsymbol{\lambda}_h$.

Return moment parameters $\boldsymbol{\eta} = \boldsymbol{\eta}(\boldsymbol{\lambda})$.

$$\pi(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\theta}) \prod_{i=1}^{n} l_i(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{2.9}$$

and the approximating distribution (2.5):

$$q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i=1}^{n} \frac{f_i(\boldsymbol{\theta})}{C_i}, \quad f_i(\boldsymbol{\theta}) = \exp\left(-\frac{1}{2}\boldsymbol{\theta}^t \mathbf{Q}_i \boldsymbol{\theta} + \mathbf{r}_i^t \boldsymbol{\theta}\right) \tag{2.10}$$

where again for simplicity, we have assumed that the true prior is Gaussian, and needs not be approximated in the course of the algorithm.

As above, the update of site i adjusts C_i, $\boldsymbol{r}_i$ and $\boldsymbol{Q}_i$ so as to minimise the Kullback-Leibler divergence between the Gaussian approximation q and a normalised version of the hybrid distribution. Simple calculations (see e.g. Seeger, 2005) lead to the following expressions for the update of C_i:

$$\log(C_i) = \log(Z_h) - \Psi(\boldsymbol{r}, \boldsymbol{Q}) + \Psi(\boldsymbol{r}_{-i}, \boldsymbol{Q}_{-i}) \tag{2.11}$$

where Z_h is the normalising constant of the hybrid, as defined in (2.8), $\boldsymbol{r}$, $\boldsymbol{Q}$ (resp. $\boldsymbol{r}_{-i}$, $\boldsymbol{Q}_{-i}$) are the natural parameters of the current Gaussian approximation q (resp. of $q/f_i \propto \prod_{j \neq i} f_j$) and $\Psi(\boldsymbol{r}, \boldsymbol{Q})$ is the log-normalising constant of an unnormalised Gaussian density:

$$\Psi(\boldsymbol{r}, \boldsymbol{Q}) = \log\left\{\int \exp\left(-\frac{1}{2}\mathbf{x}^{\mathbf{t}}\mathbf{Q}\boldsymbol{\theta} + \mathbf{r}^{\mathbf{t}}\boldsymbol{\theta}\right) d\boldsymbol{\theta}\right\} = -\frac{1}{2}\log|\boldsymbol{Q}/2\pi| + \frac{1}{2}\boldsymbol{r}^t \boldsymbol{Q} \boldsymbol{r}.$$

For each site update, one calculates C_i as defined in (2.11). Then, at the end of the algorithm, one may return the following quantity

$$\sum_{i=1}^{n} \log(C_i) + \Psi(\boldsymbol{r}, \boldsymbol{Q}) - \Psi(\boldsymbol{r}_0, \boldsymbol{Q}_0)$$

as an approximation to the logarithm of the evidence.

3 EP-ABC: Adapting EP to likelihood-free settings

3.1 Basic principle

As explained in the introduction, our objective is to approximate the following ABC posterior

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{\int p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}) \mathbb{1}_{\left\{\|y_i - y_i^\star\| \leq \varepsilon\right\}} dy_i\right\} \tag{3.1}$$

which corresponds to a particular factorisation of the likelihood,

$$p(\boldsymbol{y}^\star|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i^\star|y_{1:i-1}^\star, \boldsymbol{\theta}). \tag{3.2}$$

Note that, in full generality, the $y_i^\star$ may be any type of "chunk" of the observation vector $\boldsymbol{y}^\star$, i.e. the random variables $y_i^\star$ may have a different dimension, or more generally different supports. For simplicity, we assume that the prior $p(\boldsymbol{\theta})$ is Gaussian, with natural parameters $\boldsymbol{Q}_0$ and $\boldsymbol{r}_0$.

Algorithm 2 Computing the moments of the hybrid distribution in the likelihood-free setting, basic algorithm.
Inputs: ϵ, $\boldsymbol{y}^\star$, i, and the moment parameters $\boldsymbol{\mu}_{-i}$, $\boldsymbol{\Sigma}_{-i}$ of the Gaussian pseudo-prior q_{-i}.

1. Draw M variates $\boldsymbol{\theta}^{[m]}$ from a $N(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$ distribution.
2. For each $\boldsymbol{\theta}^{[m]}$, draw $y_i^{[m]} \sim p(y_i | y_{1:i-1}^\star, \boldsymbol{\theta}^{[m]})$.
3. Compute the empirical moments

$$M_{acc} = \sum_{m=1}^{M} \mathbb{1}_{\left\{\|y_i^{[m]} - y_i^\star\| \leq \varepsilon\right\}}, \quad \widehat{\boldsymbol{\mu}}_h = \frac{\sum_{m=1}^{M} \boldsymbol{\theta}^{[m]} \mathbb{1}_{\left\{\|y_i^{[m]} - y_i^\star\| \leq \varepsilon\right\}}}{M_{acc}} \tag{3.3}$$

$$\widehat{\boldsymbol{\Sigma}}_h = \frac{\sum_{m=1}^{M} \boldsymbol{\theta}^{[m]} \left\{\boldsymbol{\theta}^{[m]}\right\}^t \mathbb{1}_{\left\{\|y_i^{[m]} - y_i^\star\| \leq \varepsilon\right\}}}{M_{acc}} - \widehat{\boldsymbol{\mu}}_h \widehat{\boldsymbol{\mu}}_h^t. \tag{3.4}$$

Return $\widehat{Z}_h = M_{acc}/M$, $\widehat{\boldsymbol{\mu}}_h$ and $\widehat{\boldsymbol{\Sigma}}_h$.

A standard ABC algorithm is not well suited to approximate (3.1), because the prior probability that the n constraints $\|y_i - y_i^\star\| \leq \epsilon$ hold simultaneously is exponentially small with respect to n. Yet, this ABC posterior is appealing precisely for the same reason: it forces all the components of $\boldsymbol{y}$ to be close to those of $\boldsymbol{y}^\star$, and does not rely on an arbitrary choice of summary statistics.

One may interpret (3.1) as an artificial posterior distribution, which decomposes into a prior times n likelihood contributions l_i, as in (2.4), with

$$l_i(\boldsymbol{\theta}) = \left\{ \int p(y_i | y_{1:i-1}^\star, \boldsymbol{\theta}) \mathbb{1}_{\left\{\|y_i - y_i^\star\| \leq \varepsilon\right\}} \, dy_i \right\}.$$

We have seen that the feasibility of the EP algorithm is determined by the tractability of the following operation: to compute the two first moments of a pseudo-posterior, made of a Gaussian prior $N_d(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$, times a single likelihood contribution $l_i(\boldsymbol{\theta})$. This immediately suggests the following EP-ABC algorithm. We use the EP algorithm, as described in Algorithm 1, and where the moments of such a pseudo-posterior are computed as described in Algorithm 2, that is, as Monte Carlo estimates, based on simulated pairs $(\boldsymbol{\theta}^{[m]}, y_i^{[m]})$, where $\boldsymbol{\theta}^{[m]} \sim N_d(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$, and $y_i^{[m]} | \boldsymbol{\theta}^{[m]} \sim p(y_i | y_{1:i-1}^\star, \boldsymbol{\theta}^{[m]})$.

Since EP-ABC integrates one data-point at a time, it does not suffer from a curse of dimensionality with respect to n: the rejection rate of Algorithm 2 corresponds to a single constraint $\|y_i - y_i^\star\| \leq \epsilon$, not n of them, and is therefore likely to be tolerably small even for small windows ϵ.

The only requirement of EP-ABC is that the factorisation of the likelihood, (3.2), is chosen in such a way that simulating from the model, i.e. $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{\theta})$ can be decomposed into a sequence of steps, where one samples from $p(y_i | y_{1:i-1}, \boldsymbol{\theta})$, for $i = 1, \ldots, n$. We shall see in our examples section, see Section 5, that several important applications of likelihood-free inference fulfil this requirement in a trivial way. We shall also discuss in Section 6 how other likelihood-free situations may be accommodated by the EP-ABC approach.

3.2 Numerical stability

EP-ABC is a stochastic version of EP, a deterministic algorithm, hence some care must be taken to ensure numerical stability. We describe here three strategies towards this aim.

First, to ensure that the stochastic error introduced by each site update does not vary too much in the course of the algorithm, we adapt dynamically M, the number of simulated points, as follows. For a given site update, we sample repetitively M_0 pairs $(\boldsymbol{\theta}^{[m]}, y_i^{[m]})$, as described in Algorithm 2, until the total number of accepted points exceeds some threshold $M_{\min}$. Then we compute the moments (3.3) and (3.4) based on all the accepted pairs.

Second, EP-ABC computes a great deal of Monte Carlo estimates, based on IID (independent and identically distributed) samples, part of which are Gaussian. Thus, it seems worthwhile to implement variance reduction techniques that are specific to the Gaussian distribution. After some investigation, we recommend the following quasi-Monte Carlo approach. We generate a Halton sequence $\boldsymbol{\xi}^{[m]}$ of dimension d, which is a low discrepancy sequence in $[0,1]^d$, and take

$$\boldsymbol{\theta}^{[m]} = \boldsymbol{\mu}_{-i} + \boldsymbol{L}_{-i} \Phi^{-1}\left(\boldsymbol{\xi}^{[m]}\right), \quad \boldsymbol{L}_{-i} \boldsymbol{L}_{-i}^t = \boldsymbol{\Sigma}_{-i}$$

where $\boldsymbol{\mu}_{-i}$, $\boldsymbol{\Sigma}_{-i}$ are the moment parameters corresponding to the natural parameters $\boldsymbol{r}_{-i}$, $\boldsymbol{Q}_{-i}$, $\boldsymbol{L}_{-i}$ is the Cholesky lower triangle of $\boldsymbol{\Sigma}_{-i}$, and Φ^{-1} returns a vector that contains the $N(0,1)$ inverse distribution function of each component of the input vector. We recall briefly that a low discrepancy sequence in $[0,1]^d$ is a deterministic sequence that spreads more evenly over the hyper-cube $[0,1]^d$ than a sample from the Uniform distribution would; we refer the readers to e.g. Chap. 3 of Gentle (2003) for a definition of Halton and other low discrepancy sequences, and the theory of quasi-Monte Carlo. Rigorously speaking, this quasi-Monte Carlo version of EP-ABC is a hybrid between Monte Carlo and quasi-Monte Carlo, because the $y_i^{[m]}$ are still generated using standard Monte Carlo. However, we do observe a dramatic improvement when using this quasi-Monte Carlo approach. An additional advantage is that one may save some computational time by generating once and for all a very large sequence of $\Phi^{-1}\left(\boldsymbol{\xi}^{[m]}\right)$ vectors, and store it in memory for all subsequent runs of EP-ABC.

The third measure we may take is to slow down the progression of the algorithm such as to increase stability, by conservatively updating the parameters of the approximation in Step 3 of Algorithm 1, that is, $\boldsymbol{\lambda}_i \leftarrow \alpha(\boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{-i}) + (1-\alpha)\boldsymbol{\lambda}_i$. Standard EP is the special case with $\alpha = 1$. Updates of this type are suggested in Minka (2004).

In our experiments, we found that the two first strategies improve performance very significantly, and that the third strategy is sometimes useful, for example in our reaction time example, see Section 5.4.

3.3 Evidence approximation

In this section, we normalise the ABC posterior (3.1) as follows:

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) = \frac{1}{p_\epsilon(\boldsymbol{y}^\star)} p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int p(y_i|y_{1:i-1}^\star, \boldsymbol{\theta}) \frac{\mathbb{1}_{\left\{\|y_i - y_i^\star\| \leq \varepsilon\right\}}}{v_i(\epsilon)} \, dy_i \right\}, \tag{3.5}$$

where $v_i(\epsilon)$ is the normalising constant of the Uniform distribution with respect to the ball of centre $y_i^\star$, radius ϵ, and norm $\|\cdot\|$. For the Euclidean norm, and assuming that the y_i's have the same dimension d_y, one has: $v_i(\epsilon) = v_i(1)\epsilon^{d_y}$, with $v_i(1) = \pi^{d_y/2}/\Gamma(d_y/2+1)$; e.g. $v_i(1) = 2$ if $d_y = 1$, $v_i(1) = \pi$ if $d_y = 2$.

Wilkinson (2008) shows that a standard ABC posterior such as (1.1) can be interpreted as the posterior distribution of a new model, where the summary statistics are corrupted with a uniformly-distributed noise (assuming these summary statistics are sufficient). The expression above indicates that this interpretation also holds for this type of ABC posterior, except that the artificial model is now such all the random variables y_i are corrupted with noise (conditional on $y_{1:i-1}^\star$).

The expression above also raises an important point regarding the approximation of the evidence. In (3.5), the normalising constant $p_\epsilon(\boldsymbol{y}^\star)$ is the evidence of the corrupted model, which converges to the evidence $p(\boldsymbol{y}^\star)$ of the actual model as $\epsilon \to 0$. On the other hand, EP-ABC targets (3.1), and, in particular, see Section 2.4, produces an EP approximation of its normalising constant, which is $p_\epsilon(\boldsymbol{y}^\star)\prod_{i=1}^n v_i(\epsilon)$. Thus, one needs to divide this EP approximation by $\prod_{i=1}^n v_i(\epsilon)$ in order recover an approximation of $p_\epsilon(\boldsymbol{y}^\star)$. We found in our simulations that, when properly normalised as we have just described, the approximation of the evidence provided by EP-ABC is particularly accurate, see Section 5. In contrast, standard ABC based on summary statistics cannot provide an approximation of the evidence, as explained in the Introduction.

4 Speeding up EP-ABC in the IID case

Typically, the main computational bottleneck of EP-ABC, or other types of ABC algorithms, is simulating data-points from the model. In this section, we explain how these simulations may be recycled throughout the iterations in the IID (independent and identically distributed) case, so as to significantly reduce the overall computational cost of EP-ABC.

Our recycling scheme is based on a a straightforward importance sampling strategy. Consider an IID model, with likelihood $p(\boldsymbol{y}^\star|\boldsymbol{\theta}) = \prod_{i=1}^n p(y_i^\star|\boldsymbol{\theta})$. Assume that, for a certain site i, pairs $(\boldsymbol{\theta}^{[m]}, y^{[m]})$ are generated from $q_{-i}(\boldsymbol{\theta})p(y|\boldsymbol{\theta})$, as described in Algorithm 2. We have removed the subscript i in both $y^{[m]}$ and $p(y|\boldsymbol{\theta})$, to highlight the fact that the generative process of the data-points is the same for all the sites. The next update, for site $i+1$, requires computing moments with respect to $q_{-(i+1)}(\boldsymbol{\theta})p(y|\boldsymbol{\theta})\mathbb{1}\left\{\|y - y_{i+1}^\star\| \leq \epsilon\right\}$. Thus, we may recycle the simulations of the previous site by assigning to each pair $(\boldsymbol{\theta}^{[m]}, y^{[m]})$ the importance sampling weight:

$$w_{i+1}^{[m]} = \frac{q_{-(i+1)}\left(\boldsymbol{\theta}^{[m]}\right)}{q_{-i}\left(\boldsymbol{\theta}^{[m]}\right)} \times \mathbb{1}\left\{\|y^{[m]} - y_{i+1}^\star\| \leq \epsilon\right\}$$

and compute the corresponding weighted averages.

Obviously, this step may also be applied to the subsequent sites, $i+2, i+3, \ldots$, until one reaches a stage when the weighted sample is too degenerated. When this happens, "fresh" simulations may be generated from the current

Algorithm 3 Computing the moments of the hybrid distribution in the likelihood-free setting, recycling scheme for IID models.

Inputs: i, ϵ, current weighted sample $(\boldsymbol{\theta}^{[m]}, y^{[m]})_{m=1}^M$, moment parameters $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ (resp. $\boldsymbol{\mu}_{-i}$ and $\boldsymbol{\Sigma}_{-i}$) that correspond to the site where data were re-generated for the last time (resp. that correspond to the Gaussian approximation $\prod_{j\neq i} q_j(\boldsymbol{\theta})$).

1. Compute the importance sampling weights

$$w_i^{[m]} = \frac{N(\boldsymbol{\theta}^{[m]}; \boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})}{N(\boldsymbol{\theta}^{[m]}; \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})} \times \mathbb{1}_{\left\{\|y^{[m]} - y_i^\star\| \leq \epsilon\right\}}$$

   where $N(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the Gaussian $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ probability density evaluated at $\boldsymbol{\theta}$, and the effective sample size:

$$\text{ESS} = \frac{\left(\sum_{m=1}^M w_i^{[m]}\right)^2}{\sum_{m=1}^M \left(w_i^{[m]}\right)^2}.$$

2. If ESS $<$ $\text{ESS}_{\min}$, replace $(\boldsymbol{\theta}^{[m]}, y^{[m]})_{m=1}^M$ by M IID draws from $q_{-i}(\boldsymbol{\theta})p(y|\boldsymbol{\theta})$, set $w_i^{[m]} = \mathbb{1}_{\left\{\|y^{[m]} - y_i^\star\| \leq \epsilon\right\}}$, and $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_{-i}$, $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{-i}$.

3. Compute the following importance sampling estimates:

$$\widehat{Z}_h = \frac{1}{M}\sum_{m=1}^M w_i^{[m]}, \quad \widehat{\boldsymbol{\mu}}_h = \frac{\sum_{m=1}^M w_i^{[m]} \boldsymbol{\theta}^{[m]}}{\widehat{Z}_h}$$

   and

$$\widehat{\boldsymbol{\Sigma}}_h = \frac{\sum_{m=1}^M w_i^{[m]} \boldsymbol{\theta}^{[m]} \left\{\boldsymbol{\theta}^{[m]}\right\}^t}{\widehat{Z}_h} - \widehat{\boldsymbol{\mu}}_h \widehat{\boldsymbol{\mu}}_h^t.$$

Return $(\boldsymbol{\theta}^{[m]}, y^{[m]})_{m=1}^M$, $\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\Sigma}}$, $\widehat{Z}_h$, $\widehat{\boldsymbol{\mu}}_h$ and $\widehat{\boldsymbol{\Sigma}}_h$.

site. Algorithm 3 describes more precisely this recycling strategy. To detect weight degeneracy, we use the standard ESS (Effective Sample Size) criterion of Kong et al. (1994): we regenerate when the ESS is smaller that some threshold $\text{ESS}_{\min}$.

The slower the EP approximation evolves, the less often regenerating the data-points is necessary, so that as the approximation gradually stabilises, we do not need to draw any new samples any more. Since EP slows down rapidly during the first two passes, most of the computational effort will be devoted to the early phase, and additional passes through the data will come essentially for free.

In non-IID cases several options are still available. For some models the data may come in blocks, each block made up of IID data-points (think for example of a linear model with discrete predictors). We can apply the strategy outlined above in a block-wise manner (see the reaction times example, section 5.4). In other models there may be an easy transformation of the samples for data-point i such that they become samples for data-point $j \neq i$, or one may be able to reuse part of the simulation.

5 Case studies

5.1 General methodology

In each scenario, we apply the following approach. In a first step, we run the EP-ABC algorithm. We may run the algorithm several times, to evaluate the Monte Carlo variability of the output, and we may also run for different values of ε, in order to assess the sensitivity to this approximation parameter.

In a second step, we run alternative algorithms, that is, either an exact (but typically expensive) MCMC algorithm, or an ABC algorithm, based on some set of summary statistics. The ABC algorithm we implement is a Gaussian random walk version of the MCMC-ABC algorithm of Marjoram et al. (2003). This algorithm targets a standard ABC approximation, i.e. (1.1), that corresponds to a single constraint $\{\|\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\boldsymbol{y}^\star)\| \leq \epsilon\}$, for some vector of summary statistics $\boldsymbol{s}$, and some ϵ; the specific choices of $\boldsymbol{s}$ and ϵ are discussed for each application. We calibrate the tuning parameters of these MCMC algorithms using the information provided by the first step: we use as a starting point for the MCMC chain the

expectation of the approximated posterior distribution provided by the EP-ABC algorithm, random walk scales are taken to be some fraction of the square root of the approximated posterior variances, and so on. This makes our comparisons particularly unfavourable to EP-ABC. Despite this, we find consistently that the EP-ABC algorithm is faster by several orders of magnitude, and leads to smaller overall approximation errors. We report computational loads both in terms of CPU time (e.g. 30 seconds) and in terms of the number of simulations of replicate data-points y_i. The latter should be typically the bottleneck of the computation.

All the computations were performed on a standard desktop PC in Matlab; programs are available from the first author's web page.

5.2 First example: Alpha-stable Models

Alpha-stable distributions are useful in areas (e.g. Finance) concerned with noise terms that may be skewed, may have heavy tails and an infinite variance. A univariate alpha-stable distribution does not admit a close-form expression for its density, buy may be specified through its characteristic function

$$\Phi_X(t) = \begin{cases} \exp\left[i\delta t - \gamma^\alpha |t|^\alpha \left\{1 + i\beta \left(\tan\frac{\pi\alpha}{2}\right) \mathrm{sgn}(t) \left(|\gamma t|^{1-\alpha} - 1\right)\right\}\right] & \alpha \neq 1 \\ \exp\left[i\delta t - \gamma |t| \left\{1 + i\beta \frac{2}{\pi} \mathrm{sgn}(t) \log|\gamma t|\right\}\right] & \alpha = 1 \end{cases}$$

where α determines the tails, $0 < \alpha \leq 2$, β determines skewness, $-1 < \beta < 1$, and $\gamma > 0$ and δ are respectively scale and location parameters; see Nolan (2012, Chap. 1) for a general introduction to stable distributions.

Peters et al. (2010) consider a model of n i.i.d. observations y_i, $i = 1, \ldots, n$ from a univariate alpha-stable distribution, and propose to use the ABC approach to infer the parameters. Likelihood-free inference is appealing in this context, because sampling from an alpha-stable distribution is fast (using e.g. the algorithm of Chambers et al., 1976), while computing its density is cumbersome.

Trying EP-ABC on this example is particularly interesting for the following reasons: (a) Peters et al. (2010) show that choosing a reasonable set of summary statistics for this problem is difficult, and that several natural choices lead to strong biases; and (b) since alpha-stable distributions are very heavy-tailed, the posterior distribution may be heavy-tailed as well, which seems a challenging problem for a method based on a Gaussian approximation such as EP-ABC.

Our data consist of $n = 1264$ rescaled log-returns, $y_t = 100 * \log(z_t/z_{t-1})$, computed from daily exchange rates z_t of AUD (Australian Dollar) recorded in GBP (British Pound) between 1 January 2005 and 1 December 2010. (These data are publicly available on the Bank of England's web-site.) We take $\boldsymbol{\theta} = \left(\Phi^{-1}(\alpha/2), \Phi^{-1}\left((\beta+1)/2\right), \log\gamma, \delta\right)$ where Φ is the $N(0,1)$ cumulative distribution function, and we set the prior to $N\left(0_4, \mathrm{diag}(1,1,10,10)\right)$. Note however that our results are expressed in terms of the initial parametrisation α, β, γ and δ; i.e. for each parameter we report the approximate marginal posterior distribution obtained through the appropriate variable transform of the Gaussian approximation produced by EP-ABC. We run the EP-ABC algorithm (recycling version, as model is IID, see Section 4), with $\epsilon = 0.1$, $M = 8 \times 10^6$, $\mathrm{ESS}_{\min} = 2 \times 10^4$, and $\|\cdot\|$ set to the Euclidian norm in $\mathbb{R}$ (i.e. the n constraints in (3.1) simplify to $|y_i - y_i^\star| \leq \varepsilon$). Variations over ten runs are negligible. Average CPU time for one run is 39 minutes, and average number of simulated data-points over the course of the algorithm, is 4×10^8.

We first compare these results with the output of an exact random-walk Hastings-Metropolis algorithm, which relies on the evaluation of an alpha-stable probability density function for each data-points (using numerical integration). Because of this, this algorithm is very expensive. We ran the exact algorithm for about 60 hours (2×10^5 iterations). One sees in Figure 5.1 that the difference between EP-ABC and the exact algorithm are negligible. Results from the exact algorithm must be taken with some caution, as more iterations would have been required according to standard MCMC convergence checks; autocorrelation time is higher than 100 in certain dimensions.

We then compare these results with those obtained by MCMC-ABC, for the set of summary statistics which performs best among those discussed by Peters et al. (2010, see S_1 in Section 3.1). We run 2×10^7 iterations of this sampler, which leads to about 50 times more simulations from an univariate alpha-stable distribution than in the EP-ABC runs above. Through pilot runs, we decided to set $\epsilon = 0.03$, which seems to be as small as possible, subject to having a reasonable acceptance rate (2×10^{-3}) for this computational budget. In Figure 5.1, we see that the posterior output from this MCMC-ABC exercise is much more biased than EP-ABC. As explained in the previous section, we have set the starting point of the MCMC-ABC chain to the posterior mode. If initialised from some other point, the sampler typically takes a much longer time to reach convergence, because the acceptance rate is significantly lower in regions far from the posterior mode.

Finally, we also use EP-ABC, with the same settings as above, e.g. $\epsilon = 0.1$, in order to approximate the evidence of the model above (-1385.8) and two alternative models, namely a symmetric alpha-stable model, where β is set to 0 (-1383.8), and a Student model (-1383.6), with 3 parameters (scale γ, position δ, degrees of freedom ν, and a Gaussian prior $N(0_3, \mathrm{diag}(10,10,10))$ for $\boldsymbol{\theta} = (\log\nu, \gamma, \delta)$). (Standard deviation over repeated runs is below 0.1.) One sees that there

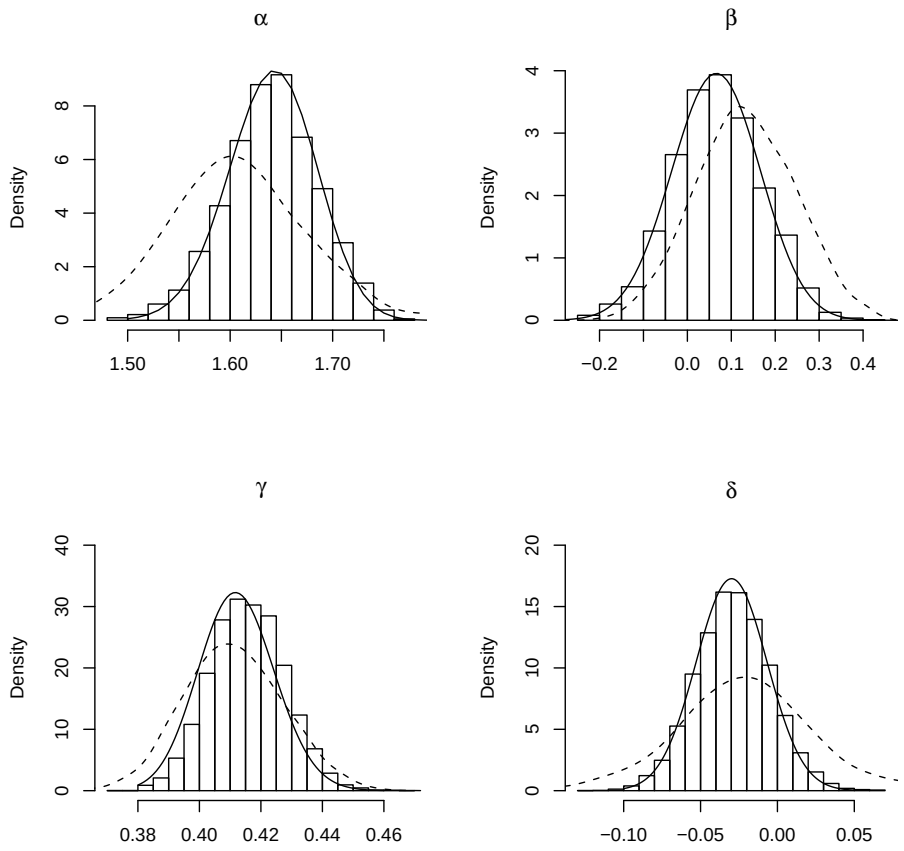


Fig. 5.1: Marginal posterior distributions of α, β, γ and δ for alpha-stable model: MCMC output from the exact algorithm (histograms), approximate posteriors provided by first run of EP-ABC (solid line), kernel density estimates computed from MCMC-ABC sample based on summary statistic proposed by Peters et al. (2010) (dashed line).

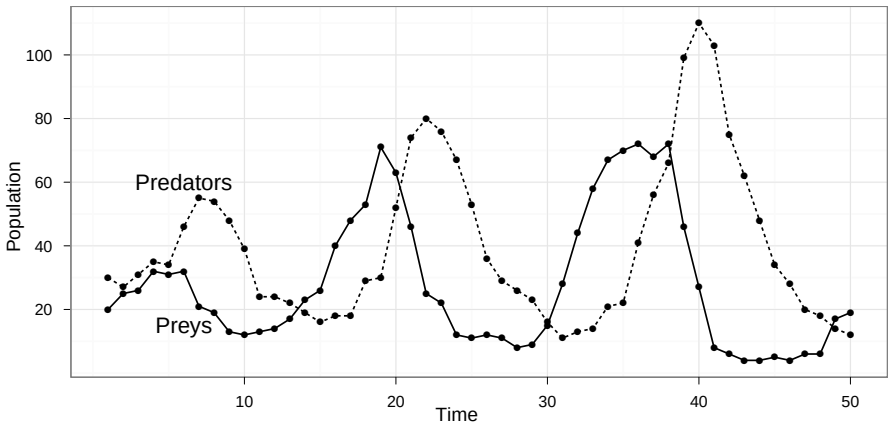


Fig. 5.2: Lokta-Volterra example: simulated dataset

no strong evidence of skewness in the data, and that the Student distribution and a symmetric alpha-stable distribution seem to fit equally well the data. We obtained the same value (-1383.6) for the evidence of the Student model when using the generalised harmonic mean estimator (Gelfand and Dey, 1994) based on a very long chain of an exact MCMC algorithm. For the two alpha-stable models, this approach proved to be too expensive to allow for a reliable comparison.

5.3 Second example: Lotka-Volterra models

The stochastic Lotka-Volterra process describes the evolution of two species Y_1 (prey) and Y_2 (predator) through the reaction equations:

$$\begin{aligned} Y_1 &\xrightarrow{r_1} 2Y_1 \\ Y_1 + Y_2 &\xrightarrow{r_2} 2Y_2 \\ Y_2 &\xrightarrow{r_3} \emptyset \end{aligned}$$

This chemical notation means that, in an interval $[t, t+dt]$, the probability that one prey is replaced by two preys is $r_1 d_t$, and so on. Typically, the observed data $\boldsymbol{y}^\star = (y_1, \ldots, y_n)$ are made of n vectors $y_i^\star = (y_{i,1}^\star, y_{i,2}^\star)$ in $\mathbb{N}^2$, which correspond to the population levels at integer times. We take $\boldsymbol{\theta} = (\log r_1, \log r_2, \log r_3)$. This model is Markov, $p(y_i^\star | y_{1:i-1}^\star, \boldsymbol{\theta}) = p(y_i^\star | y_{i-1}^\star, \boldsymbol{\theta})$ for $i > 1$, and one can efficiently simulate from $p(y_i^\star | y_{i-1}^\star, \boldsymbol{\theta})$ using Gillespie (1977)'s algorithm. On the other hand, the density $p(y_i^\star | y_{i-1}^\star, \boldsymbol{\theta})$ is intractable. This makes this model a clear candidate both for ABC, as noted by Toni et al. (2009), and for EP-ABC. Boys et al. (2008) show that MCMC remains feasible for this model, but in certain scenarios the proposed schemes are particularly inefficient, as noted also by Holenstein (2009, Chap. 4).

Following the aforementioned papers, we consider a simulated dataset, corresponding to rates $r_1 = 0.4$, $r_2 = 0.01$, $r_3 = 0.3$, initial population values $y_{0,1}^\star = 20$, $y_{0,2}^\star = 30$ and $n = 50$; see Figure 5.2. Since the observed data are integer-valued, we use the supremum norm in (3.1), and an integer-valued ϵ; this is equivalent to imposing simultaneously the $2n$ constraints $\left|y_{i,1} - y_{i,1}^\star\right| \leq \epsilon$ and $\left|y_{i,2} - y_{i,2}^\star\right| \leq \epsilon$ in the ABC posterior.

First, we run EP-ABC (standard version) with $M_{\min} = 4000$, and for both $\epsilon = 3$ and $\epsilon = 1$. We find that a single pass over the data is sufficient to reach convergence. For $\epsilon = 3$ (resp. $\epsilon = 1$), CPU time for each run is 2.5 minutes (resp. 25 minutes), and number of simulated transitions $p(y_i | y_{i-1}^\star, \boldsymbol{\theta})$ is about 10^7 (resp. 9×10^7); marginal posteriors obtained through EP-ABC are reported in Figure 5.3.

When applying ABC to this model, Toni et al. (2009) uses as a pseudo-distance between the actual data $\boldsymbol{y}^\star$ and the simulated data $\boldsymbol{y}$ the sum of of squared errors. In Wilkinson (2008)'s perspective discussed in Section 4, this is equivalent to considering a state-space model where the latent process is the Lokta-Volterra process described above, and the observation process is the same process, but corrupted with Gaussian noise. Thus, instead of a standard ABC algorithm, one may use a MCMC sampler specifically designed for state-space models in order to simulate from the ABC approximation of Toni et al. (2009). Following Holenstein (2009, Chap. 4), we consider a Gaussian random walk version of the marginal PMCMC sampler. This algorithm is a particular instance of the state of the art PMCMC framework of Andrieu et al. (2010), which is based on the idea of approximating the marginal likelihood of the data by running a particle filter of size N at each iteration of the MCMC sampler.

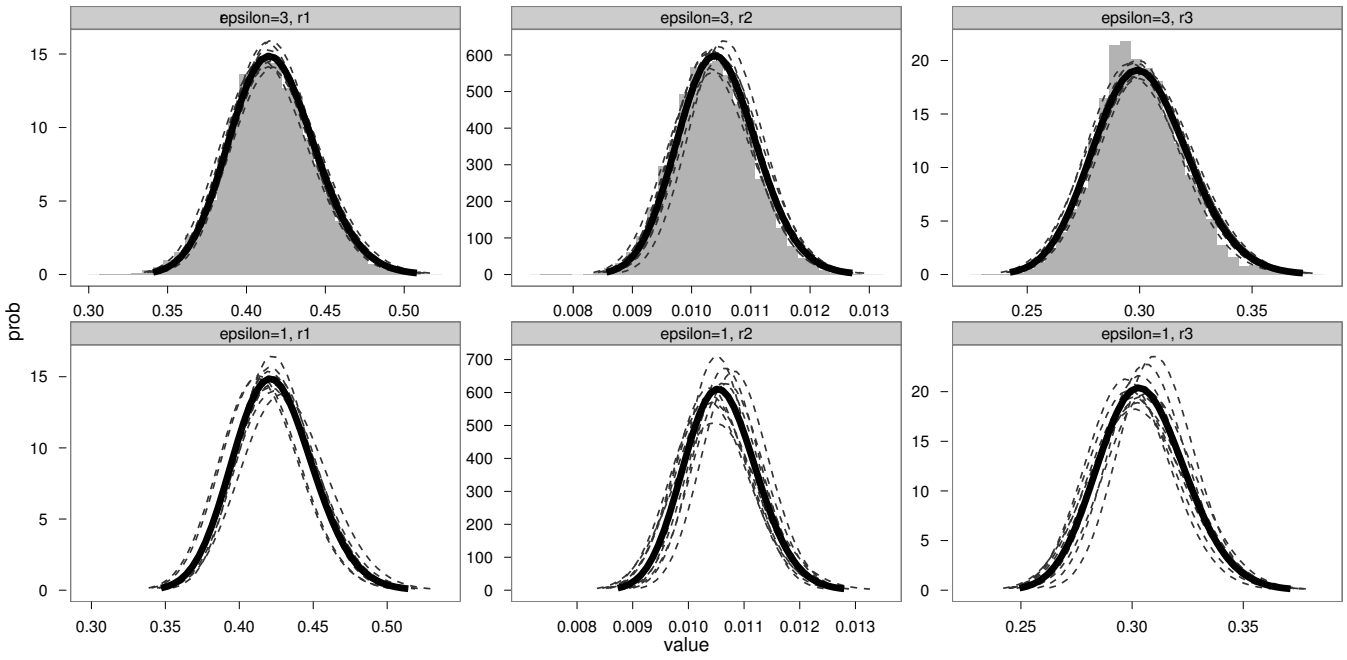


Fig. 5.3: Lokta-Volterra example: marginal posterior densities of rates r_1, r_2, r_3, obtained from PMCMC algorithm (histograms), and from ABC-EP, for $\epsilon = 3$ (top) and $\epsilon = 1$ (bottom); PMCMC results for $\epsilon = 1$ could not be obtained in a reasonable time. The solid lines correspond to the average over 10 runs of the moment parameters of the Gaussian approximation, while the dashed lines correspond to the 10 different runs.

In Figure 5.3, we report the posterior output obtained from this sampler, run for about 2×10^5 iterations and $N = 1000$ particles (2 days in CPU time, 10^{10} simulated transitions $p(y_i|y^\star_{i-1}, \boldsymbol{\theta})$), with random walk scales set to obtain a 0.25 acceptance rate. These plots correspond to $\epsilon = 3$, and a state-space model with an uniformly distributed observation noise. In Figure 5.3, one detects practically no difference between PMCMC and EP-ABC with $\epsilon = 3$ (black lines), although the CPU time of the latter was about 1500 smaller.

The difference between the two EP-ABC approximations (corresponding to $\epsilon = 1$ and $\epsilon = 3$) is a bit more noticeable. Presumably, the EP-ABC approximation corresponding to $\epsilon = 1$ is slightly closer to the true posterior. We did not manage however to obtain reliable results from our PMCMC sampler and $\epsilon = 1$ in a reasonable time.

5.4 Third example: Race models of reaction times

Reaction time models seek to describe the decision behaviour of (human or animal) subjects in a choice task (Luce, 1991; Meyer et al., 1988; Ratcliff, 1978). In the typical experiment, subjects view a stimulus, and must choose an appropriate response. For example, the stimulus might be a set of moving points, and the subject must decide whether the points move to the left or to the right.

Assuming that the subject may choose between k alternatives, one observes independent pairs, $y_i = (d_i, r_i)$, where $d_i \in \{1, \ldots, k\}$ is the chosen alternative, and $r_i \geq 0$ is the measured reaction time. For convenience, we drop for now the index i in order to describe the random distribution of the pair (d, r).

Reaction time models assume that the brain processes information progressively, and that a decision is reached when a sufficient amount of information has been accumulated. In the model we use here (a variant of typical models found in e.g. Ratcliff and McKoon, 2008; Bogacz et al., 2007) k parallel integrators represent the evidence $e_1(t), \ldots, e_k(t)$ in favour of each of the k alternatives. The model is illustrated on Figure 5.4. The first accumulator to reach its boundary b_j wins the race and determines which response the subject will make. Each accumulator undergoes a Wiener process with drift:

$$\tau de_j(t) = m_j dt + dW_t^j$$

where the m_j's are the drift parameters, the W_t^j's are k independent Wiener processes; and τ is a fixed time scale, $\tau = 5ms$. The measured reaction time is corrupted by a uniformly-distributed noise r_{nd}, representing the "non-decisional

time" (Ratcliff and McKoon, 2008), i.e. the time the subject needs to execute the decision (prepare a motor command, press an answer key, ...). This model is summarised by the following equations:

$$\begin{aligned} r &= r_d + r_{nd}, \quad r_{nd} \sim \mathrm{U}[a,b], \\ r_d &= \min_j \inf_t \{t : e_j(t) = b_j\}, \\ d &= \arg\min_j \inf_t \{t : e_j(t) = b_j\}. \end{aligned}$$

(We fix a and b to $a = 100ms$, $b = 200ms$, credible values from Ratcliff and McKoon (2008))

The model above captures the essential ideas of reaction time modelling, but it remains too basic for experimental data. We now describe several important extensions. First, a better fit is obtained if the boundaries are allowed to vary randomly from trial to trial (as in Ratcliff, 1978): we assume that $b_j = c_j + \tau$, where $\tau \sim N(0, e^s)$, and s is a parameter to be estimated. Second, a mechanism is needed to ensure that the reaction times cannot be too large: we assume that if no boundary has been reached after 1 second, information accumulation stops and the highest accumulator determines the decision. Finally, one needs to account for lapses (Wichmann and Hill, 2001): on certain trials, subjects simply fail to pay attention to the stimuli and respond more or less at random. We account for this phenomenon by having a lapse probability of 5%. In case a lapse occurs, r_d becomes uniformly distributed between 0 and 800 ms and the response is chosen between the alternatives with equal probability. Clearly, this generalised model remains amenable to simulation, although the corresponding likelihood is intractable.

We apply this model to data from an unpublished experiment by M. Maertens (personal communication to Simon Barthelmé). The dataset is made of 1860 observations, obtained from a single human subject, which had to choose between $k = 2$ alternatives: "signal absent" (no light increment was presented), or "signal present" (a light increment was presented), under 15 different experimental conditions: 3 different locations on the screen, and 5 different contrast values. Following common practice in this field, trials with very high or very low reaction times (top and bottom 5%) were excluded from the dataset, because they have a high chance of being outliers (fast guesses, keyboard errors or inattention). The data are shown on Figure 5.5.

From the description above, one sees that five parameters, (m_1, m_2, c_1, c_2, s), are required to describe the random behaviour of a single pair (r_i, d_i), when $k = 2$. To account for the heterogeneity introduced by the varying experimental conditions, we assume that the 2 accumulation rates, m_1, m_2 vary across the 15 experimental conditions, while the 3 parameters related to the boundaries, c_1, c_2 and s, are shared across conditions. The parameter $\boldsymbol{\theta}$ is therefore 33-dimensional.

We note that this model would present a challenge for inference even if the likelihood function was available. It is difficult to assign priors to the parameters, because they do not have a clear physical interpretation, and available prior knowledge (e.g., that reaction times will normally be less than 1 second) does not map easily unto them. Moreover, the model is subject in certain cases to weak identifiability problems. For instance, if one response dominates the dataset, there is little information available beyond the fact that one drift rate is much higher than the other (or one threshold much lower than the other, or both).

We re-parametrised the positive parameters c_1, c_2 as $c_1 = e^{\lambda}$, $c_2 = e^{\lambda+\delta}$, and assigned a $\mathcal{N}(0,1)$ prior to λ, δ, and s. Taking a $N(0, 5^2)$ prior for these 3 quantities led to similar results. Some experimentation suggested that the drift rates could be constrained to lie between -0.1 and 0.1, because values outside of this interval seem to yield improbable reaction times (too short or too long). We assigned a $[-0.1, 0.1]$ uniform prior for the 30 drift rates, and applied the appropriate transform, i.e. $x \to 0.1 + 5\Phi^{-1}(x)$, in order to obtain a $N(0,1)$ prior for the transformed parameters.

After a few unsuccessful attempts, we believe that this application is out of reach of normal ABC techniques. The main difficulty is the choice of the summary statistics. For instance, if one takes basic summaries (e.g. quartiles) of the distribution of reactions times, under each of the 15 experimental conditions, one ends with a very large vector $\boldsymbol{s}$ of summary statistics. Due to the inherent curse of dimensionality of standard ABC (Blum, 2010), sampling enough datasets, of size 1860, which are sufficiently close (in terms of the pseudo-distance $\|\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\boldsymbol{y}^\star)\|$) would require enormous computational effort. Obviously, taking a much smaller set of summary statistics would on the other land lead to too poor an approximation.

Some adjustments are needed for ABC-EP to work on this problem. First, notice that within a condition the datapoints are IID so that the posterior distribution factorises over IID "blocks". We can therefore employ the recycling technique described in Section 4 to save simulation time, by going through the likelihood sites block-by-block. Second, since datapoints take values in $\{1,2\} \times \mathbb{R}^+$, we adopt the following set of ABC constraints: $\mathbb{1}\{d_i = d_i^\star\}\,\mathbb{1}\{|\log r_i - \log r_i^\star| \leq \epsilon\}$, where $y_i^\star = (d_i^\star, r_i^\star)$ denotes as usual the actual datapoints. Third, we apply the following two variance-reduction techniques. One stems from the fact that each site likelihood does not depend on all the 33 parameters but on a subset of size 5. In that case, using simple linear algebra, one can see that is possible to update only the marginal distribution of the EP approximation with respect to these 5 parameters; see the Appendix for details. The second is a simple Rao-Blackwellsiation scheme

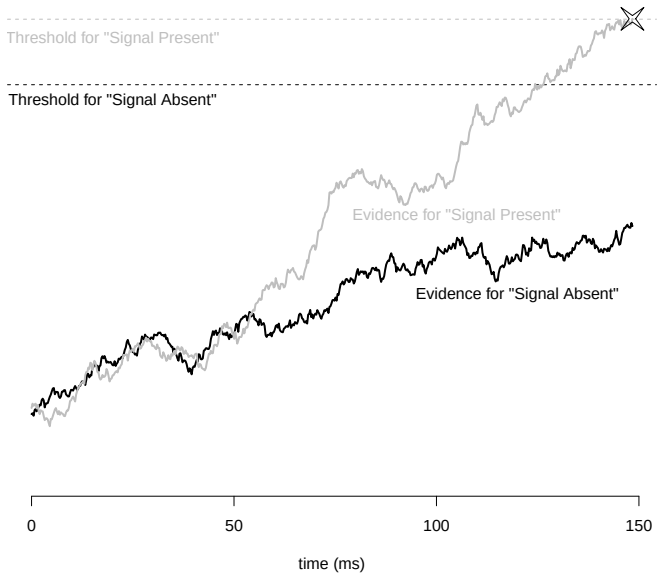
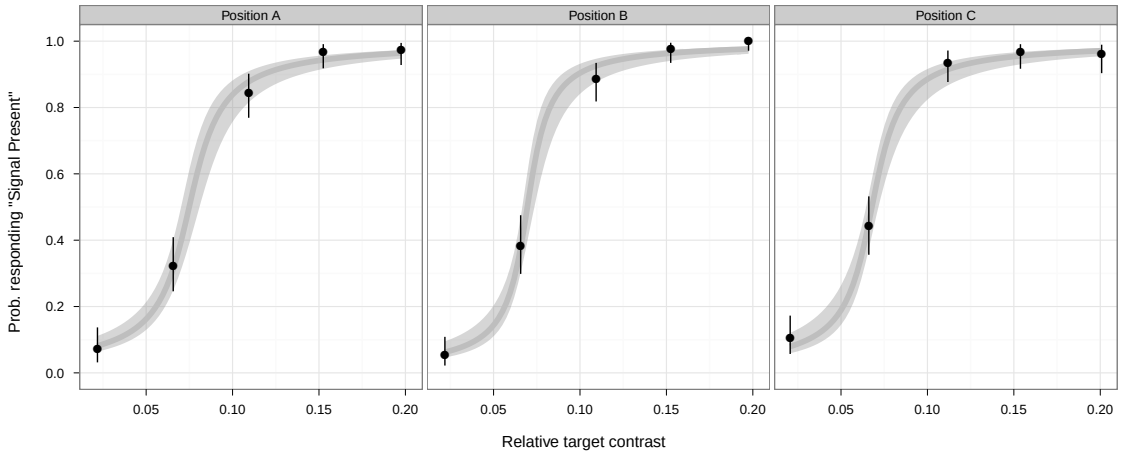
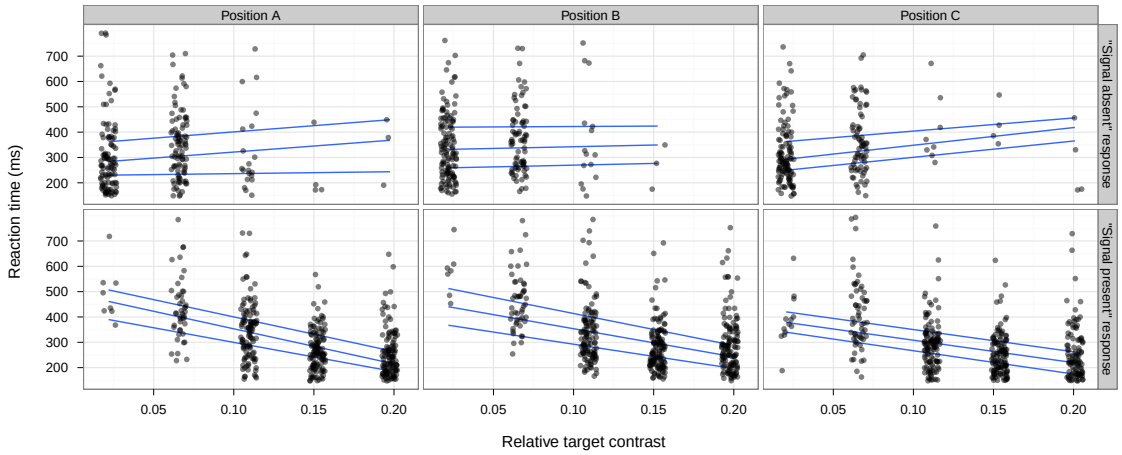


Fig. 5.4: A model of reaction time in a choice task. The subject has to choose between k responses (here, "Signal present" and "Signal absent") and information about each accumulates over time in the form of evidence in favour of one and the other. Because of noise in the neural system "evidence" follows a random walk over time. A decision is reached when the evidence for one option reaches a threshold (dashed lines). The decision time in this example is denoted by the star: here the subject decides for 'B' after about 150ms. The fact that the thresholds are different for "Signal Present" and "Signal Absent" capture decisional bias: in general, for the same level of information, the subject favours option "Signal Absent".



(a) Probability of answering "Signal present" as a function of relative target contrast in a detection experiment, at three different positions of the target (data from one subject). Filled dots represent raw data, the grey curves are the result of fitting a binomial GLM with Cauchit link function. The light grey band is a 95% confidence band. As expected in such experiments, the probability of detecting the target increases with relative target contrast.



(b) Reaction time distributions conditional on target contrast, target position, and response. The semi-transparent dots represent reaction times for individual trials. Horizontal jitter was added to aid visualisation. The lines represent linear quantile regressions for the 30%, 50% and 70% quantiles.

Fig. 5.5: Choice (**a**) and reaction time (**b**) data in a detection experiment by Maertens.

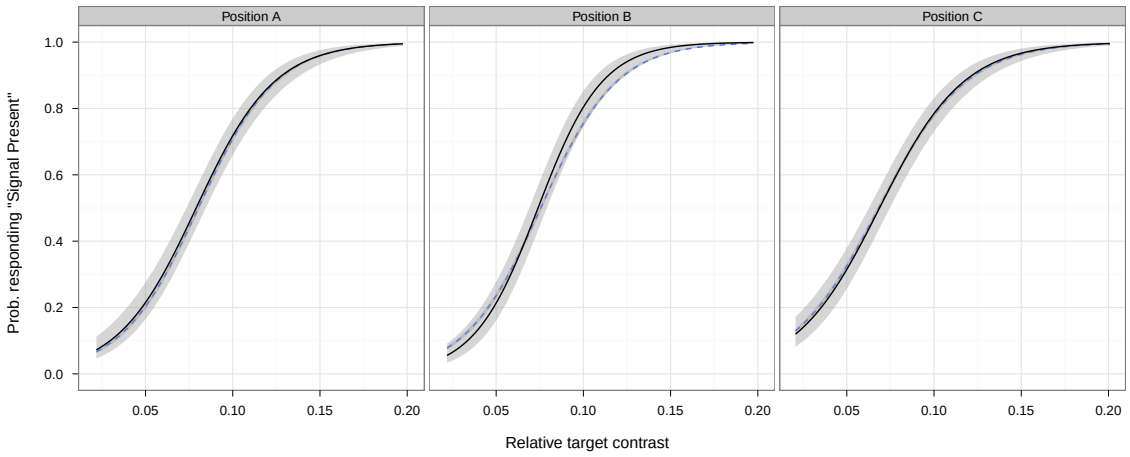


Fig. 5.6: Probability of answering "Signal Present" as a function of contrast: data versus (approximate) posterior predictive distribution. We sampled the posterior predictive distribution and summarised it using a binomial GLM with Cauchit link, in the same way as we summarised the data (see (a) in Figure 5.5). The data are represented as a continuous line, the predictive distribution as dotted. The posterior predictive distribution is very close to the data. The shaded area corresponds to a 95% confidence interval for the fit to the data.

that uses the fact that the non-decisional component r_{nd} is uniformly distributed, and may therefore be marginalised out when computing the EP update.

We report below the results obtained from ABC-EP with $\epsilon = 0.16$, $M = 3 \times 10^3$, $\alpha = 0.4$ (see end of Section 3.2), and 3 complete passes over the data; CPU time was about 40 minutes. Results for smaller values of ϵ, e.g. $\epsilon = 0.1$, were mostly similar, but required a larger CPU time.

Since we could not compare the results to those of a standard ABC algorithm, we asses the results through posterior predictive checking. For each of the 15 experimental conditions, we generate 5,000 samples from the predictive density, and compare the simulated data with the real, as follows.

The marginal distribution of responses can be summarised by regressing the probability of response on stimulus contrast, separately for each stimulus position (as on Figure 5.5), and using a binomial generalized linear model (with Cauchit link function). Figure 5.6 compares data and simulations, and shows that the predictive distribution successfully captures the marginal distribution of responses.

To characterise the distribution of reaction times, we look at means and inter-quantile intervals, conditional on the response and the experimental conditions. The results are presented on Figure 5.7. The predictive distributions capture the location and scale of the reaction time distributions quite well, at least for those conditions with enough data. Such results seem to indicate that, at the very least, the ABC-EP approximate posterior corresponds to a high-probability region of the true posterior.

6 Discussion

In its current presentation, EP-ABC seems to have two limitations. First, it assumes a Gaussian prior; and second, it relies on a particular factorisation of the likelihood, which makes it possible to simulate sequentially the datapoints.

The first limitation is mostly a technicality. EP-ABC, like any EP algorithm, may accommodate a non-Gaussian prior, by just treating the prior as an additional site. We have not described this generalisation in the paper, because we find more expedient in practice to simply reparametrise the model so as to make the prior Gaussian, as we did implicitly in our numerical examples. Since the only input required by EP-ABC from the model is an algorithm that samples a data-point, for a given parameter value, applying a particular one-to-one transform to the parameter is trivial.

The second limitation requires more discussion. In this paper, we focused our attention on important applications of likelihood-free inference where the likelihood is trivial to factorise; either because the datapoints are independent, or Markov. But ABC-EP is not limited to these two situations. First, quite a few time series models may be simulated sequentially, even if they are not Markov. For instance, one may apply straightforwardly EP-ABC to a GARCH-stable model (e.g. Liu and Brorsen, 1995; Mittnik et al., 2000; Menn and Rachev, 2005), which is a GARCH model (Bollerslev, 1986) with an alpha stable-distributed innovation. Second, one may obtain a simple factorisation of the likelihood by incorporating latent variables into the vector $\boldsymbol{\theta}$ of unknown parameters. The most obvious application of this idea is

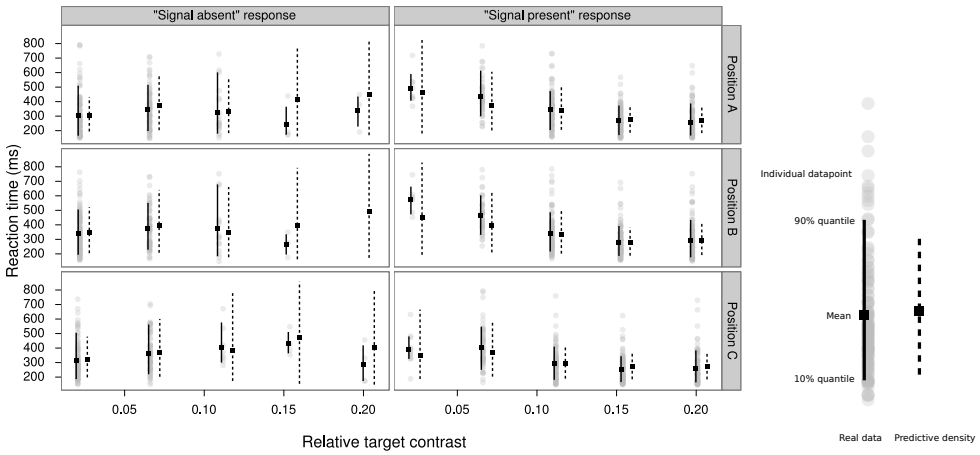


Fig. 5.7: Reaction time distributions conditional on decision: data versus posterior predictive distributions. The reaction times conditioned on contrast, position and response are shown as grey dots and summarised via mean and 10-90% inter-quantile range (continuous lines). The posterior predictive distributions computed from samples are summarised and shown with an offset to the right (dotted lines). The conditional densities are well captured, given sufficient data.

hidden Markov models, where the observations are conditionally independent, given the fixed parameters and the hidden process; see e.g. Dean et al. (2011), Calvet and Czellar (2011) for applications of likelihood free inference to this class of models. In such a situation, however, one needs to ensure that EP-ABC provides stable results, despite the possibly large dimension of $\boldsymbol{\theta}$. The marginal update outlined in the Appendix may be helpful in this context. Third, it might be possible in certain situations, such as repeated experiments, or in the presence of mixed effects (in the spirit of our vision example), to use EP-ABC in order to break the posterior down in several "pieces", and treat each piece through standard ABC techniques. At all rates, we believe that such extensions of EP-ABC are exciting avenues for future research.

Finally, one may always criticise EP-ABC as being based on EP, a machine learning method which shows good performance empirically in many situations, but which currently lacks a proper mathematical justification, both in terms of convergence of the algorithm, and in terms of assessing the approximation error. Work in this direction has started recently (Titterington, 2011), but these preliminary results do not address the issue of the stability of the algorithm when each site update introduces a stochastic error, as in EP-ABC. This is another important direction for future research, and EP-ABC may well require a more technical convergence study, because of its stochastic nature.

As of now, we would like to reply to this criticism by two remarks. First, even if EP-ABC delivers accurate results directly, the user is free to use EP-ABC as a first, quick step, in order to to calibrate a second, more expensive step based on a standard ABC approximation. Second, and this may well be the main message of this paper: it seems quite absurd to reject an EP-based approach, if the only alternative is an ABC approach based on summary statistics, which introduces a bias which seems both larger (according to our numerical examples) and more arbitrary, in the sense that in real-world applications, one has little intuition and even less mathematical guidance on to why $p(\boldsymbol{\theta}|\boldsymbol{s}(\boldsymbol{y}))$ should be close to $p(\boldsymbol{\theta}|\boldsymbol{y})$ for a given set of summary statistics $\boldsymbol{s}$. Not to mention that introducing summary statistics prevents from computing the evidence of the model, as we have already explained. In fact, we argue that this dependence on summary statistics is currently the main limitation of the ABC approach, and that it is essential that this issue is addressed in future research in likelihood-free inference, whether through EP or by other means.

Acknowledgements

The authors thank Pierre Jacob for insightful comments, and M. Maertens for providing the data used in the third example. The first author acknowledges support from the BMBF (Foerderkennzeichen 01GQ1001B). The second author acknowledges support from the "BigMC" ANR grant ANR-008-BLAN-0218 of the French Ministry of Research.

Appendix: marginal EP updates

In example 2 we make use of the fact that some sites only depend on a subset of the parameters to obtain more stable updates. This strategy is potentially very important for the extension of ABC-EP to Hidden Markov models suggested in

the discussion. We list below some results for multivariate Gaussian families that are essential in deriving these special EP updates.

We generalise the problem slightly to computing the moments of a hybrid $h(\boldsymbol{\theta}) \propto f(\mathbf{A}\boldsymbol{\theta})\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, the product of a multivariate Gaussian density and a likelihood which is a function of $\mathbf{A}\boldsymbol{\theta}$, where $\mathbf{A}$ is a matrix of dimension $k \times m$, $m < k$. When $\mathbf{A}$ is a sub-matrix of the identity matrix we have the special case of a likelihood which only depends on a subset of the parameters.

For the normalisation constant, we have that:

$$\begin{aligned} \int f(\mathbf{A}\boldsymbol{\theta})\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\,\mathrm{d}\boldsymbol{\theta} &= \int f(\mathbf{z}) \left(\int_{\{\boldsymbol{\theta}:\mathbf{A}\boldsymbol{\theta}=z\}} \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\,\mathrm{d}\boldsymbol{\theta} \right) \mathrm{d}\mathbf{z} \\ &= \int f(\mathbf{z})\mathcal{N}\left(\mathbf{z}, \mathbf{A}\boldsymbol{\mu}_0, \mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\right)\mathrm{d}\mathbf{z} \end{aligned} \tag{6.1}$$

where we have defined a new variable $\mathbf{z} = \mathbf{A}\boldsymbol{\theta}$. Let us denote by γ the above integration constant. We can now derive an expression for the mean of the hybrid:

$$\begin{aligned} \gamma^{-1} \int \boldsymbol{\theta} f(\mathbf{A}\boldsymbol{\theta})\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\,\mathrm{d}\boldsymbol{\theta} &= \gamma^{-1} \int f(\mathbf{z}) \left(\int_{\{\boldsymbol{\theta}:\mathbf{A}\boldsymbol{\theta}=z\}} \boldsymbol{\theta}\mathcal{N}(\boldsymbol{\theta}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\,\mathrm{d}\boldsymbol{\theta} \right) \mathrm{d}\mathbf{z} \\ &= \gamma^{-1} \int f(\mathbf{z})\mathcal{N}\left(\mathbf{z}, \mathbf{A}\boldsymbol{\mu}_0, \mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\right) E(\boldsymbol{\theta}|\mathbf{z})\mathrm{d}\mathbf{z} \end{aligned} \tag{6.2}$$

where $E(\boldsymbol{\theta}|\mathbf{z})$ is the conditional expectation of $\boldsymbol{\theta}$ given $\mathbf{z}$. For multivariate Gaussians $E(\boldsymbol{\theta}|\mathbf{z})$ is a linear function of $\mathbf{z}$:

$$E(\boldsymbol{\theta}|\mathbf{z}) = \mathbf{V}\mathbf{z} + \mathbf{b} \tag{6.3}$$

where $\mathbf{V} = \boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\left(\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\right)^{-1}$ and $\mathbf{b} = \boldsymbol{\mu}_0 - \boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\left(\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\right)^{-1}\mathbf{A}\boldsymbol{\mu}_0$.
We inject (6.3) into (6.2):

$$\begin{aligned} E_h(\boldsymbol{\theta}) &= \gamma^{-1} \int f(\mathbf{z})\mathcal{N}\left(\mathbf{z}, \mathbf{A}\boldsymbol{\mu}_0, \mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\right)(\mathbf{V}\mathbf{z} + \mathbf{b})\,\mathrm{d}\mathbf{z} \\ &= \mathbf{V}E_h(\mathbf{z}) + \mathbf{b} \end{aligned}$$

where E_h denotes expectation over the hybrid density. A similar calculation yields an expression for the covariance:

$$Cov_h(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\left(\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{A}^{\mathbf{t}}\right)^{-1}\mathbf{A}\boldsymbol{\Sigma}_0 + \mathbf{V}Cov_h(\mathbf{z})\,\mathbf{V}^{\mathbf{t}} \tag{6.4}$$

These three results yield computational savings and increased stability, because the moments of the hybrid distribution over $\boldsymbol{\theta}$ can be obtained from the moments of the marginal hybrid over $\mathbf{z}$, which has lower dimensionality.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Statist. Soc. B*, 72(3):269–342.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Bishop, C. (2006). *Pattern recognition and machine learning*. Springer New York.

Blum, M. G. B. (2010). Approximate Bayesian Computation: A Nonparametric Perspective. *J. Am. Statist. Assoc.*, 105(491):1178–1187.

Bogacz, R., Usher, M., Zhang, J., and McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485):1655–1670.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31(3):307–327.

Boys, R., Wilkinson, D., and Kirkwood, T. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statist. Comput.*, 18(2):125–135.

Brascamp, J. W., van Ee, R., Noest, A. J., Jacobs, R. H., and van den Berg, A. V. (2006). The time course of binocular rivalry reveals a fundamental role of noise. *Journal of vision*, 6(11):1244–1256.

Calvet, L. and Czellar, V. (2011). State-observation sampling and the econometrics of learning models. *Arxiv preprint arXiv:1105.4519*.

Chambers, J., Mallows, C., and Stuck, B. (1976). A method for simulating stable random variables. *J. Am. Statist. Assoc.*, 71:340–344.

Dean, T., Singh, S., Jasra, A., and Peters, G. (2011). Parameter estimation for hidden markov models with intractable likelihoods. *Arxiv preprint arXiv:1103.5399*.

Gelfand, A. and Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *J. R. Statist. Soc. B*, 56(3):501–514.

Gentle, J. (2003). *Random number generation and Monte Carlo methods*. Springer-Verlag.

Gerstner, W. and Kistler, W. M. (2002). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 1 edition.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Holenstein, R. (2009). *Particle Markov Chain Monte Carlo*. PhD thesis, University of British Columbia.

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputation and Bfayesian missing data problems. *J. Am. Statist. Assoc.*, 89:278–288.

Liu, S. and Brorsen, B. (1995). Maximum likelihood estimation of a GARCH-stable model. *J. Econometrics*, 10(3):273–285.

Luce, D. R. (1991). *Response Times: Their Role in Inferring Elementary Mental Organization (Oxford Psychology Series)*. Oxford University Press, USA.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov Chain Monte Carlo without Likelihoods. 100(26):15324–15328.

Menn, C. and Rachev, S. (2005). A GARCH option pricing model with ialpha-stable innovations. *European journal of operational research*, 163(1):201–209.

Meyer, D. E., Osman, A. M., Irwin, D. E., and Yantis, S. (1988). Modern mental chronometry. *Biological psychology*, 26(1-3):3–67.

Minka, T. (2001). Expectation Propagation for approximate Bayesian inference. *Proceedings of Uncertainty in Artificial Intelligence*, 17:362–369.

Minka, T. (2004). Power EP. Technical report, Dep. Statistics, Carnegie Mellon University, Pittsburgh, PA.

Mittnik, S., Paolella, M., and Rachev, S. (2000). Diagnosing and treating the fat tails in financial returns data. *Journal of Empirical Finance*, 7(3-4):389–416.

Nolan, J. P. (2012). *Stable Distributions - Models for Heavy Tailed Data*. Birkhauser, Boston. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

Nuthmann, A., Smith, T. J., Engbert, R., and Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, 117(2):382–405.

Peters, G., Sisson, S., and Fan, Y. (2010). Likelihood-free Bayesian inference for alpha-stable models. *Comput. Stat. Data Anal.*, (in press).

Pritchard, J., Seielstad, M., Perez-Lezaun, A., and Feldman, M. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, (85):59–108.

Ratcliff, R. and McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922.

Robert, C., Marin, J.-M., and Pillai, N. S. (2011). Why approximate Bayesian computational (ABC) methods cannot handle model choice problems. *Arxiv preprint arXiv:1101.5091*.

Seeger, M. (2005). Expectation Propagation for Exponential Families. Technical report, Univ. California Berkeley.

Titterington, M. (2011). The EM algorithm, variational approximations, and expectation propagation for mixtures. In Mengersen, K., Robert, C., and Titterington, M., editors, *Mixtures: Estimation and Applications (Chap. 1)*, volume 896. Wiley.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Wang, B. and Titterington, D. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 373–380.

Wichmann, F. A. and Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception and Psychophysics*, 63(8):1293–1313.

Wilkinson, R. (2008). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Arxiv preprint arXiv:0811.3355*.